Using Summaries in Document Retrieval

Mark Wasson

LexisNexis, a Division of Reed Elsevier, plc
9443 Springboro Pike
Miamisburg, OH, USA 45342
mark.wasson@lexis-nexis.com

Paper ID: P##### No mention of this in call for papers

Keywords: information retrieval, summary search, highly relevant documents

Contact Author: Mark Wasson

Under consideration for other conferences (specify)? No

Abstract

This paper examines the role that summaries can play in document retrieval. Thirty searches are applied to full-text and summaries only in large document collections, and the results are evaluated using two difference evaluation scopes. The results support the view that those customer segments who want smaller answer sets focused on highly relevant documents benefit from limiting their searches to summaries. On the other hand, those customer segments who wish to retrieve all references to some topic should continue to search full-text.

Using Summaries in Document Retrieval

Author Name, see cover page

AuthorEmployer
Address
Location
e-mail

Abstract

This paper examines the role that summaries can play in document retrieval. Thirty searches are applied to full-text and summaries only in large document collections, and the results are evaluated using two difference evaluation scopes. The results support the view that those customer segments who want smaller answer sets focused on highly relevant documents benefit from limiting their searches to summaries. On the other hand, those customer segments who wish to retrieve all references to some topic should continue to search full-text.

1 Introduction

The goal that motivated the creation of Searchable LEAD in news documents in the AuthorEmployer collection was to provide *some* customer segments with a tool that helps them focus their retrieval results on a limited number of *highly relevant documents*, where a highly relevant document with respect to some query is a document that is substantially about the query topic.

A good, general purpose document summary should capture the major topics presented in a document. Presumably if we can capture major topics in summaries, then a search that is restricted to summaries should do a better job of limiting retrieval results to highly relevant documents about those topics.

To support this, we created Searchable LEAD, a process that identifies and labels the leading sentences or paragraphs of news documents as a separate searchable LEAD field. A customer's query, e.g., BUSH AND GORE, could easily be limited to the LEAD, e.g., LEAD(BUSH AND GORE), or HEADLINE and LEAD combination, e.g., HLEAD(BUSH AND GORE). Through three separate experiments, the value of leading text as a general purpose summary for news documents has been verified. This paper describes a fourth experiment that investigates whether and how searches limited to this type of summary benefit the targeted customers.

In many information retrieval experiments, a single user perspective, i.e., a single answer key, is used to evaluate the results. If that perspective matches that of the targeted customer set, the evaluation is meaningful. However, different customer segments perform information seeking tasks with different goals and perspectives in mind, even when they are interested in the same topic. Just as potential search tool enhancements are not one-size-fits-all, a one-size-fits-all answer key should not be used to determine the value of a search aid for two sets of customers with fundamentally different goals. In this experiment, the results of each query were evaluated using two different user perspectives, highly relevant references only and all references. Through this approach we were able to determine whether Searchable LEAD satisfied the goal that motivated its creation.

2 Defining a Summary for News Articles

For purposes of this investigation, the leading text of news documents is used as a basis for creating document summaries – specifically the definition of Searchable LEAD found in Author (1998).

Brandow et al. (1995) compared summaries they created using *tf-idf*-based sentence extraction to fixed amounts of leading text – approximately

60, 150 and 250 words long, in three separate trials – generated using a slightly modified version of our production Searchable LEAD text processing software. In that effort, Searchable LEAD-based extracts were judged to be acceptable as summaries for general news articles 92% of the time. This compared favorably to the 74% reported for those summaries created through sentence extraction. However, that test was limited to only 250 news articles.

Author (1998) reported on a larger scale version of this evaluation, although in that work Searchable LEAD was used as-is. Searchable LEAD-based extracts resulted in an average compression ratio of 13% in that test. Compression ratios generally ranged between about 5-20% for most documents, depending on document length; with Searchable LEAD, the number of leading sentences and paragraphs included in the leading text field was linked to document length. For a shorter document, the Searchable LEAD might consist of only a single sentence. For long documents, Searchable LEAD might consist of the first three paragraphs or more of the document.

The Searchable LEAD-based extracts were evaluated on their acceptability as summaries in more than 2,727 documents. For the 1,951 general news articles in that test corpus, Searchable LEADs were judged to be acceptable as summaries 94.1% of the time, a result that is not appreciably different from that reported by Brandow et al. (1995), especially when seven newsbrief type documents are excluded from their results. For the other types of documents in the corpus, including lists, newsbriefs and transcripts, acceptability rates were somewhat to substantially lower, as Table 1 shows.

Document Type	**Number of Documents**	**Acceptability Rate**
General News	1,951	94.1%
Lists	86	12.8%
Newsbriefs	191	24.6%
Transcripts	499	70.3%

Table 1. LEAD as Summary Acceptability Rates for Document Types

Zhou (1999) reported the results of an experiment where Searchable LEADs were compared to summaries created by two internal prototype and three commercially available sentence extraction summary generators in a document relevance judgment task, where evaluators used each summary to determine the corresponding document's relevance to a topic. The result of that evaluation showed that the top five systems, including Searchable LEAD, statistically tied in this task.

3 Related Work

In addition to evaluating the value of both Searchable LEAD-based and sentence extraction-based extracts for their value as general summaries, Brandow et al. (1995) also reported on the results of a limited experiment examining the differences between summary-only versus full-text searching. In tests involving twelve Boolean queries applied to a corpus of about 20,000 documents extracted from the AuthorEmployer NEWS library. They reported that average precision increased from an average of 37% for searches applied to full-text to 45% for searches applied to sentence extraction-based extracts and 47% for searches applied to leading text-based extracts. This was more than offset by large drops in *relative recall*, 100% for full-text compared to 56% for sentence extraction-based extracts and 58% for leading text-based extracts. Relative recall assumes that the full-text queries achieved 100% recall; due to limited resources on the project, there was no attempt to determine actual recall rates.

In addition to its limited scale, there were two key problems with this evaluation. First, although Brandow et al. (1995) correctly reported that Searchable LEAD was introduced to enhance search precision, Searchable LEAD also targeted only a subset of our customer segments, specifically those customers who wanted to retrieve only highly relevant documents (in AuthorEmployer-internal jargon, we refer to these as *on-point* or *major reference* documents). This point was not mentioned in Brandow et al. (1995), nor was it reflected in their search evaluation. Second, the convenience of using relative recall notwithstanding, this approach to measuring recall will generally magnify the difference in recall that should be expected when comparing full-text and summary-only search results.

Sumita & Iida (1997) tested both leading text-based extracts and *tf-idf*-based sentence extracts of up to three sentences in an experiment involving 10 queries and 600 Japanese language news articles. They reported that limiting searches to such summaries both improved the effectiveness for retrieving highly relevant documents, but also helped exclude other relevant documents with lower levels of relevance.

Sakai & Sparck Jones (2001) examined the value of summaries for general information retrieval and a pseudo-relevance feedback model, in their case using 30 queries applied to a nearly-39,000 document corpus derived from the TREC collection. The Okapi Basic Search System was used. Precision evaluation focused on both the top 1000 and top 10 relevance ranked documents retrieved. The authors concluded that a summary-only search may be as effective as full-text for precision-oriented searching of highly relevant documents. Incorporating both summaries and full-text documents into their pseudo-relevance feedback model was significantly more effective than using summaries only.

4 User Evaluation Scopes

Most information retrieval experiments calculate recall, precision and the corresponding f-measure from a single *evaluation perspective* or *evaluation scope*. All documents are judged to be relevant or irrelevant with respect to that one scope. However, commercial information services now report that they handle millions of searches a day for their customers. It is not reasonable to assume that all of the people using those services have the same perspective on relevance, and yet that is often how we evaluate new search aids and features.

Our customers employ a variety of search strategies, depending on their topics, information interests, and the point they are at in their information seeking task. At one end, we see some customers just starting out on an information seeking task, where they typically are looking for a few highly relevant documents to help introduce themselves to the topic. Basically they are trying to provide themselves with a good starting point. At the other extreme, we see customers in public relations, competitive intelligence or in the due diligence phase of their information seeking task. These customers often want to retrieve *all references* to the topic, even those documents that provide even the most limited or mundane information.

Although some may see this simply as the customary recall-precision trade-off, that is not the case. A document that contains a passing reference to some topic is relevant to those with the *all reference evaluation scope* (retrieval of that document is considered successful recall), but it is irrelevant to those with a h*ighly relevant reference evaluation scope* (retrieval of that document is considered a precision error). A document's relevance with respect to some customer's evaluation scope is what drives customer perceptions of the resulting recall and precision. Instead of a recall-precision trade-off, we have multiple evaluation scopes for which recall and precision are determined.

We recognize the differences in evaluation scopes in a single user over time when proposing learning systems and personalization tools that adapt retrieval or routing results to a user's changing interests (e.g., Lam et al., 1996), but we do not recognize these differences when we use single answer key evaluations. As a result, over the years, we have seen a number of *potentially* useful search enhancements dismissed not because they failed to show improvement for any targeted subset of customers, but rather because they failed to show improvement when using a single general evaluation standard (Harmon, 1991; Voorhees, 1994; Sparck Jones, 1999). Query expansion functionality such as some types of morphological or synonym expansion, for example, may produce a drop in precision that offsets any improvements to recall, but we have found that customer segments who require retrieving all references to their topic are willing to put up with a lot of irrelevant information to make sure that they see everything. Of course, those customers would still *like to have* better precision, but they *require* better recall.

This was also a problem with the limited retrieval experiment reported in Brandow et al. (1995). Although Searchable LEAD was introduced specifically to support the subset of users seeking only highly relevant documents, Brandow et al. (1995) did not make this distinction when evaluating their test of twelve Boolean queries.

For each query evaluated in the experiment reported here, two user evaluation *scopes* were created. One represented Searchable LEAD's targeted customer segment and its desire to retrieve only highly relevant documents; the other represented the due diligence customer segment, which prefers to retrieve all documents that contain information about the topic regardless of how little or how much.

5 The Experiment

5.1 Test Corpus

Searchable LEAD was tested in the AuthorEmployer NEWS library, a commercial collection of full-text news documents from several thousand sources, including newspapers, magazines, wire services, abstract services, trade journals, transcript services and other sources. The document types in this document collection reflected these sources. Date-bounded subsets of this collection were used, with date ranges varying in length from one day (typically more than 45,000 documents searched) to two years (typically more than 32 million documents searched).

5.2 Search Topics and Topic Scope

For this investigation thirty topics were selected and defined. The following are a few of the topics included in the set of topics:

- General information about Exxon Corp.
- Biographical information about Bill McCartney, founder of Promise Keepers
- Office Depot revenue and earnings information
- A specific Dallas Cowboys-Cincinnati Bengals football game
- Expensive outhouses in national parks

For each of the thirty topics, two *scope statements* were created, where a scope statement is a description of what is considered a relevant document with respect to the topic.

One scope statement, the *highly relevant reference evaluation scope*, defined what would constitute a highly relevant document. These scope statements typically combined quantitative measures with a number of specific pieces of information that must be present in a retrieved document for it to be considered highly relevant. Requiring some specific pieces of information to be present added objectivity to the evaluation process.

The second scope statement, the *all reference evaluation scope*, defined the minimum information about the topic that must be present in order to consider the document relevant from that perspective. For a named entity topic, a document relevant to the all reference scope might include as little as a single occurrence of the entity's name.

The highly relevant reference evaluation scope for the *Office Depot* query required among other things revenues, earnings (loss) information, and related per-share information. The all reference evaluation scope required at least one of the financial performance measures, with revenue typically being the one found in retrieved documents.

The highly relevant reference evaluation scope for the *Dallas Cowboys-Cincinnati Bengals football game* query required some specific game statistics, none of which were required for the all reference evaluation scope. Thus, a pre-game story concerning whether a player might play was relevant to the all reference evaluation scope but it was irrelevant to the highly relevant reference evaluation scope. After all, articles written before the game took place obviously could not include game statistics.

More than half the topics focused on named entities. This is consistent with our observations of customer search topics applied to news data, and this user behavior has also been reported elsewhere (e.g., Thompson & Dozier, 1997). One effect of this was that the recall and precision rates we would observe in this experiment were higher than what is commonly reported for Boolean search results. Because many proper names are relatively unambiguous, and because articles about some named entity almost always mention the name, some of the queries had much higher accuracy rates than might otherwise be expected, and that pulled overall average accuracy rates up somewhat. The Boolean search **EXXON**, for example, virtually assures us of 100% recall regardless of which evaluation scope is used. Although individual Exxon service stations are mentioned periodically in the news, most news articles that mention Exxon are in fact about the major oil company, ensuring fairly high precision for the all references evaluation scope.

5.3 Queries

Searchable LEAD was created to be used with a Boolean search engine. With 20% of news documents in our archives being less than 100 words long, a sizeable number of documents have one-sentence LEADs, which would be of little value to search engines that rely on term frequency.

Through our own experience and routine observations of World Wide Web searchers, most customer queries are quite short, typically one or two words or phrases, perhaps connected by one Boolean operator. Similarly short queries were created for use in this evaluation, such as the following:

- **EXXON**
- **BILL MCCARTNEY**
- **OFFICE DEPOT AND EARNINGS**
- **BENGALS AND COWBOYS**
- **NATIONAL PARK AND OUTHOUSE**

In some cases, a date restriction was explicitly added to the query. In all other cases, a most recent two-year period default date restriction was used.

There was no attempt to maximize the accuracy of the queries tested. Rather, the goal was to use queries that mimic typical user behavior in order to see how Searchable LEAD impacts typical users.

5.4 Testing

Each query was applied and evaluated in four ways, once for each evaluation scope-text scope combination:

- All reference, full-text
- All reference, LEAD only
- Highly relevant reference, full-text
- Highly relevant reference, LEAD only

The all reference/full-text combination was evaluated first. Because this combination retrieves at least all the documents retrieved by any of the other search-evaluation scope combinations, it was possible to use the results of this evaluation to create an answer key that could also be used by the other evaluations in order to ensure consistency of relevance judgments with respect to evaluation scope for all the combinations.

Each test query was applied to all of the documents in date-restricted subsets of the All News (ALLNWS) file in the AuthorEmployer NEWS library. A date restriction was used to limit the number of documents to be examined when verifying the results. In addition to applying and evaluating the query created for a given topic, additional queries were used in order to find potential recall errors, that is, relevant documents with respect to the evaluation scope of the topic that were missed by the original query. For the *Dallas Cowboys-Cincinnati Bengals football game* topic, for example, in addition to the test query **BENGALS AND COWBOYS**, other queries used to search the date range of documents in order to identify potential recall errors included the following:

- **CINCINNATI AND (COWBOYS OR FOOTBALL) AND NOT(BENGALS AND COWBOYS)**
- **DALLAS AND (BENGALS OR FOOTBALL OR OHIO) AND NOT(BENGALS AND COWBOYS)**
- **CINERGY AND NOT(BENGALS AND COWBOYS)** (name of the football field where the game was played)

All documents retrieved by such queries were examined for their degree of relevance in order to produce more accurate recall results in this test.

There was no particular attempt to match the date range exactly to a specific event, a characteristic of this test (and typical user behavior) that often contributed to the number of precision errors. For example, the *Dallas Cowboys-Cincinnati Bengals football game* occurred in the previous week, specifically three days earlier but documents retrieved from the entire week were examined. Criteria for a highly relevant reference to this game included certain game statistics. Stories written before the game could not possibly include such information, so they were counted as precision errors for the highly relevant reference evaluation scope. From a customer's perspective, our routine reverse chronological presentation of retrieved documents would have effectively hidden such errors from customers until after the desired information was obtained. For evaluation purposes, however, the entire date range was evaluated.

Full-text queries were limited to HEADLINE and BODY fields of documents. LEAD only queries were limited to the LEAD sub-field of the BODY field. Most news documents in the AuthorEmployer service also have one or more meta-data fields that may include named entity and/or topic-indicating controlled vocabulary terms, in addition to other information. Limiting queries to the HEADLINE, BODY and LEAD fields focused the evaluation on the impact of using summaries as opposed to that of using other possible editorial enhancements of the data.

5.5 Evaluation

The purpose of Searchable LEAD as a retrieval aid is to help some customer segments retrieve a highly relevant documents about some topic, and to minimize the number of irrelevant documents and documents that only contain passing references to the topic in the answer set. If Searchable LEAD works, one would expect that queries restricted to the LEAD field would result in higher precision than queries applied to the full-text would.

For the all reference evaluation scope, one would expect recall to fall when shifting from full-text to LEAD. After all, a general summary like LEAD typically only includes information on major points in the document.

The impact on recall for the highly relevant reference evaluation scope is less certain. Given that the Searchable LEAD represents an acceptable summary in only 94% of general news articles, and a lower figure in other types of documents found in the AuthorEmployer NEWS library, it is also reasonable to assume that some decline in recall would also occur with this evaluation scope. Given that relevant documents with this evaluation scope must include all the targeted information, recall errors as defined by this scope may actually eliminate information redundancy, and thus are not necessarily critical to the customer. However, the way in which basic pieces of information are presented can also be revealing, so such redundant documents may still be useful. Calculating recall in these cases thus is still worthwhile.

Recall and precision rates were calculated for each query for each evaluation scope-text scope combination. For each full-text/LEAD pair, recall and precision rates were compared to see how consistent increases and decreases were with respect to expectations.

6 Results

Thirty queries were applied first to full-text and then limited to LEAD only. The results of each approach were evaluated twice, once from the all reference evaluation scope and once from the highly relevant reference evaluation scope.

For the customer perspective that Searchable LEAD targets – the highly relevant reference evaluation scope – there was a sizeable improvement in answer set precision, which increased an average of .286, from an average of .230 to an average of .516 when the query was limited to the LEAD. As Table 2 also shows, recall decreased an average of .192 across the thirty queries. The average standard f-measure across the thirty queries increased .150, from .300 to .450.

	Full-text	**LEAD**
Avg. Recall	.785	.593
Avg. Precision	.230	.516
Avg. f-measure	.300	.450

Table 2. Averages for thirty queries using the highly relevant reference evaluation scope.

(NOTE: The f-measure listed under LEAD is lower than both the corresponding recall and precision. Keep in mind that the figures above represent averages for thirty queries. The f-measure .450 thus is not based on a recall rate of .593 and a precision rate of .516 but rather it is the average of thirty individual f-measures. This also explains f-measures provided in Tables 3 and 4.)

When evaluated from the perspective of customers who want to retrieve all references to a topic, restricting the query to the LEAD on average resulted in a substantial drop in recall, from an average of .704 to an average of .232, a drop of .472 on average. The small .082 increase in average precision rates barely provides any offsetting benefits, as Table 3 shows. Average f-measures across the thirty queries dropped .324. Customers who want to retrieve all references not surprisingly do not benefit at all from using Searchable LEAD.

	Full-text	**LEAD**
Recall	.704	.232
Precision	.777	.859
f-measure	.657	.333

Table 3. Averages for thirty queries using the all reference evaluation scope.

The trends represented by these tables were generally consistent with the results for individual queries. When using the highly relevant reference evaluation scope, precision rates and f-measures increased for 26 of the thirty queries when shifting from full-text to LEAD. When using the all reference evaluation scope, recall rates decreased or stayed steady and f-measures decreased for all thirty queries when shifting from full-text to LEAD.

For all queries tested, the number of documents retrieved when using Searchable LEAD not surprisingly was lower than when using full-text. Searchable LEAD-based answer sets on average were one fourth the size of full-text-based answer sets, 50.6 documents vs. 198.7 documents, respectively.

7 Discussion

As Table 2 and answer set size statistics show, targeted customers benefit when limiting their queries to the Searchable LEAD. As Table 3 suggests, non-targeted customers such as those who want documents with any references to the topics, clearly should not use Searchable LEAD.

Most information retrieval system evaluations do not take differing customer perspectives into account. If we were to combine the results of our all reference and highly relevant reference evaluation scopes into one general evaluation pool as is done in Table 4, we would still note a significant improvement in precision. However, based on the falling f-measure, some might conclude that summaries are simply another failed attempt to improve information retrieval with the help of natural language processing. For the average customer, that is probably a fair conclusion. However, for the customer segments that prefer to retrieve a few good highly relevant documents as they start an information seeking task, Searchable LEAD helped to produce smaller answer sets that were more focused on the highly relevant documents that those customers target.

	Full-text	**LEAD**
Avg. Recall	.745	.413
Avg. Precision	.504	.688
Avg. f-measure	.479	.392

Table 4. Averages for thirty queries combining both the highly relevant reference and all reference evaluation scopes.

As for other retrieval tasks, when creating a document categorization system, we did gain some benefits when weighting terms found in headlines and leading text in news documents a bit higher (Author, 2000), but that effect is limited to news data. A colleague investigating an internal *tf-idf*-based search engine found no benefits to putting extra emphasis on terms found in the first paragraph of news articles, but that was a rather limited test. Neither of these were evaluated from multiple user perspectives, although in the case of Author (2000) the original project goal was to identify and categorize only highly relevant documents

8 Conclusion

Customers of online services approach their information seeking tasks from many perspectives, and yet most IR evaluations are conducted from a single user perspective. Using a single user perspective is easier, but it risks hiding the potential benefits of some new feature from key customer segments who might value it. News document summaries such as those in the form of Searchable LEAD provide a way to help some customer segments retrieve smaller answer sets that are focused on highly relevant documents. But the benefits of this are only apparent when the results are evaluated from that perspective.

It is not simply a trade-off between recall and precision, but one between recall and precision with respect to the definition of relevance that that different customer segments have. An answer set of ten documents that mention but do not comment on some topic may result in 100% precision for the all reference evaluation scope but 0% precision for the highly relevant reference evaluation scope. Customer segments can and do have such widely divergent views of relevance.

The evaluation here showed that a customer seeking to retrieve a few highly relevant documents about some topic would benefit from using Searchable LEAD, retrieving smaller answer sets and a higher proportion of highly relevant documents in that answer set. A customer wanting to retrieve all documents that refer to the topic should avoid Searchable LEAD and instead continue to use full-text search.

References

Author, X. (1998) *Using Leading Text for News Summaries: Evaluation Results and Implications for Commercial Summarization Applications*. COLING-ACL '98 Conference Proceedings.

Author, X. (2000) *Large-scale Controlled Vocabulary Indexing for Named Entities*. Proceedings of the 6th Applied Natural Language Processing Conference.

Brandow, R., Mitze, K. and Rau, L. (1995) *Automatic Condensation of Electronic Publications by Sentence Selection*. Information Processing & Management, 31/5.

Harmon, D. (1991) How Effective is Suffixing? Journal of the American Society for Information Science, 42/1.

Lam, W., Mukhopadhyay, S., Mostafa, J., and Palakal, M. (1996) *Detection of Shifts in User Interests for Personalized Information Filtering*. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Sakai, T., and Spark Jones, K. (2001). *Generic Summaries for Indexing in Information Retrieval*. Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Sparck Jones, K. (1999) *What is the Role of NLP in Text Retrieval*. In "Natural Language Information Retrieval", T. Strzalkowski, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands.

Sumita, E., & Iida, H. (1997). *Information Retrieval Using a Statistical Abstraction Model*. Proceedings of the 3rd Annual Meeting of the Association for Natural Language Processing.

Thompson, P. and Dozier, C. (1997) *Name Searching and Information Retrieval*. Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing.

Voorhees, E. (1994) *Query Expansion Using Lexical Semantic Relations*. Proceedings of the 17th International ACM-SIGIR Conference on Research and Development in Information Retrieval.

Zhou, J. (1999) *Phrasal Terms in Real-World IR Applications*. In "Natural Language Information Retrieval", T. Strzalkowski, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands.